Matching Structure for Dual Learning

Hao Fei [1 2] **Shengqiong Wu** [1] **Yafeng Ren** [3] **Meishan Zhang** [4]

Abstract

Many natural language processing (NLP) tasks appear in dual forms, which are generally solved by dual learning technique that models the dualities between the coupled tasks. In this work, we propose to further enhance dual learning with structure matching that explicitly builds structural connections in between. Starting with the dual text↔text generation, we perform dually-syntactic structure co-echoing of the region of interest (`RoI`) between the task pair, together with a syntax cross-reconstruction at the decoding side. We next extend the idea to a text↔non-text setup, making alignment between the syntactic-semantic structure. Over 2*14 tasks covering 5 dual learning scenarios, the proposed structure matching method shows its significant effectiveness in enhancing existing dual learning. Our method can retrieve the key `RoIs` that are highly crucial to the task performance. Besides NLP tasks, it is also revealed that our approach has great potential in facilitating more non-text↔non-text scenarios.

1. Introduction

A good number of NLP tasks come in dual forms, such as neural machine translation (NMT) (He et al., 2016a), paraphrase generation (Ma et al., 2018), image captioning (Aneja et al., 2018) vs. text-to-image generation (van den Oord et al., 2016), text classification (Zhang et al., 2016) vs. conditioned text generation (Hu et al., 2017), semantic parsing (Gardner et al., 2018) vs. language generation (Wong & Mooney, 2007), etc. Dual learning therefore has been proposed to model the duality between the primal and dual tasks, by minimizing the gap between joint distributions of the two tasks respectively (He et al., 2016a; Xia et al., 2017; 2018). Effectively capturing the inline features between the task pair and bringing significant improvements, dual learning methods have received increasing research attention within recent years in relevant communities (Xu et al., 2018; Su et al., 2019; Cao et al., 2019; Peng & Wang, 2019; Shen & Feng, 2020).

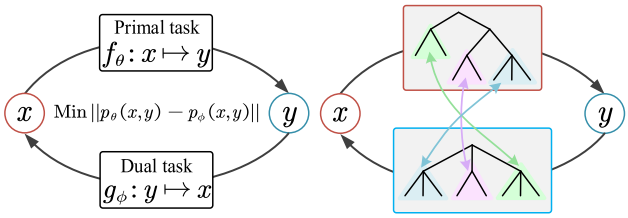


Figure 1. **Left**: dual learning framework. **Right**: dual learning with alignment of structural supervision.

We notice that the current dual learning scheme, however, fails to explicitly model the structure correspondence between two coupled tasks. The integration of structure knowledge has been extensively exploited for enhancing the feature learning in a wide range of NLP tasks (Eriguchi et al., 2016; Marcheggiani & Titov, 2017; Ponti et al., 2018; Shi et al., 2019; Sun et al., 2019; Akoury et al., 2019; Kumar et al., 2020; Bugliarello & Elliott, 2021), which offers additional bias from a lower-level perspective (e.g., syntactic or linguistic) for better task-semantic inference (Chen et al., 2019b; Fei et al., 2020a; Goyal & Durrett, 2020; Fei et al., 2021a; Wu et al., 2021). Unfortunately, the study of structure integration for dual learning has been kept unexplored. Given a pair of task, not only do they share the same input and output (in reverse), but it is often a close correspondence of the intermediate structures between them. Taking the classic NMT as example, a source sentence always shares rich syntactic structure alignments, e.g., the phrasal constituents, with the target sentence (Stahlberg et al., 2016). There are also various cases in other dual-learning scenarios that can benefit from the structure alignments, such as text-to-image, text-to-audio, etc.

To close the gap, this paper proposes matching the structure for dual learning. As shown in Figure 1, based on the vanilla

[1]School of Computing, National University of Singapore, Singapore [2]Sea-NExT Joint Lab, Singapore [3]School of Interpreting and Translation Studies, Guangdong University of Foreign Studies, China [4]Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen), China. Correspondence to: Yafeng Ren <renyafeng@whu.edu.cn>, Meishan Zhang <mason.zms@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

dual learning framework, we perform structural alignment unsupervisedly between the primal and dual tasks, bridging them with structure connections. Since the textual sentences naturally come with syntactic structures, we start with the dual modeling of text↔text for NLP, performing dually-syntactic structure matching (cf. §4). We use constituency tree as the underlying structure of each, where the phrases well represent the compositional semantics of a sentence. At the fine-grained scope, we encourage each specific region of interest (i.e., RoI) to align with the corresponding one in the opposite task as much as possible. The match measurement is conducted automatically according to the similarity. At the global scope, we perform structural cross-reconstruction, generating target text and meanwhile reconstructing the target syntactic structure.

The above idea is designed for a text↔text case. We next extend the architecture to the text↔non-text one, so as to adapt to broader application scenarios and modalities (e.g., label, image or audio). As the task of non-text side lacks explicit syntactic structure, we alternatively take its semantic structure instead, and match the syntactic structure with the text-side task, i.e., performing syntactic-semantic structure aligning (cf. §5). Correspondingly, we conduct the syntactic-semantic RoI alignment at the local scope and perform structural unilateral-reconstruction at the global scope.

Our method is verified on numbers of dual applications, including text↔text, text↔image and text↔label, where significant improvements are achieved against the vanilla dual learning. Via further evaluations we gain more findings: **1)** Our model effectively retrieves the key RoIs that are crucial to the task improvements, and strengthens the duality between dual tasks by correctly aligning the RoIs. **2)** The structural co-echoing offers rich syntactic signals for content planning in text↔text generation scenarios, leading to better diversification and grammar correctness. **3)** The success of structure matching can be extended to non-text↔non-text dual learning. **4)** The richer the structural information for the alignment, the better the improvements our method presents.

Overall, our key contributions are as follows:

• We for the first time introduce the idea of structure co-echoing for dual learning, reinforcing the structure connections between two coupled tasks.

• We study dually-syntactic structure matching for the dual text↔text generation, in which we propose a syntactic RoI alignment and a structural cross-reconstruction strategies from two different perspectives, respectively.

• We extend the structure matching architecture for the text↔non-text dual learning, by measuring the semantic-syntax structure correspondence. We further empirically explore the extendibility for non-text↔non-text applications.

• The proposed method gains improvements over 2*14 tasks spanning 5 dual learning scenarios. Further in-depth analyses and insights are shown.[1]

2. Related Work

Many a learning task in machine learning areas (e.g., NLP and computational visual) takes dual forms (Xia et al., 2017; Ye et al., 2019). The coupled tasks have the same exact input and output but in reverse. For example, an English-to-French translation task is the dual task of French-to-English translation task (He et al., 2016a); the automatic speech recognition and the text-to-speech (Tjandra et al., 2017) and the question answering vs. question generation (Sun et al., 2020) etc. The primal task and the dual task form a closed loop, generating informative feedback signals that can actually benefit both two tasks. Correspondingly, dual learning technique (He et al., 2016a) has been proposed for exploiting the duality of the task pairs. In the process, two dual tasks are jointly learned, in which the the intrinsic probabilistic connection in between are explicitly strengthened, pushing the learning process towards the right direction. Since it greatly magnifies the effectiveness of the task performances in pair without using additional annotations, the dual learning has received consistent research attention (Xu et al., 2018; Su et al., 2019; Shen & Feng, 2020). Several different paradigms of dual leaning are explored, including the unsupervised dual leaning He et al. (2016a) and the supervised dual learning (DSL) Xia et al. (2017). This work follows the line of DSL, taking it as the backbone framework. Our aim is to further reinforce the duality of DSL by encouraging each fine-grained region of interest of one task to align with the corresponding regions of interest in the opposite task as much as possible.

In NLP community, the linguistic syntax structures fundamentally describe the underlying working about how the words or phrases connect to each others, and then compose into sentence or discourse. Thus, the syntactic information, e.g., phrasal constituency tree (Aarts & Aarts, 1982) and dependency tree (Hays, 1964), are extensively integrated into wide range of downstream NLP tasks as types of external knowledge for performance enhancements (Socher et al., 2013; Nguyen & Shirai, 2015; Garmash & Monz, 2015; Marcheggiani & Titov, 2017; Zhang et al., 2018; Fei et al., 2021b). In this work, we for the first time in literature extend the success of the integration of syntactic structure information into the dual learning. Different from those existing syntactic-aware works that focus on integrating them into one singleton task process, we fuse the syntactic knowledge into both two dual tasks, then match the syntactic structures of the primal and dual tasks, and reinforce the fine-grained

[1]Source codes are available at https://github.com/scofield7419/StruMatchDL

structural correspondences (i.e., RoI) in between, by which we expect to promote the dual learning utility for the both two sides of the tasks.

3. Dual Learning Backbone

Formally, a dual learning system comprises 1) a primal task that maps $x \in \mathcal{X}$ to $y \in \mathcal{Y}$, i.e., $f_\theta : x \mapsto y$; and 2) a dual task mapping $y \in \mathcal{Y}$ to $x \in \mathcal{X}$, i.e., $g_\phi : y \mapsto x$. Each separate task has its learning object for empirical risk minimization with cross-entropy loss:

$$\mathcal{L}_\theta = \mathbb{E}_{x,y} \log p(y|x;\theta)\,, \tag{1}$$
$$\mathcal{L}_\phi = \mathbb{E}_{x,y} \log p(x|y;\phi)\,. \tag{2}$$

Let's summarize them as $\mathcal{L}_C = \mathcal{L}_\theta + \mathcal{L}_\phi$.

The primal and dual tasks should have a joint probabilistic duality (Xia et al., 2017):

$$\begin{aligned} & p_\theta(x,y) = p(x)p(y|x;\theta) \\ \simeq\ & p_\phi(x,y) = p(y)p(x|y;\phi)\,, \forall x \& y\,, \end{aligned} \tag{3}$$

where $p(x)$ and $p(y)$ are the marginal distributions which often are intractable. The dual learning targets encouraging the task pair to optimize their duality, i.e., narrowing the gap between their joint distributions:

$$\begin{aligned} \mathcal{L}_D = & || \log \hat{p}(x) + \log p(y|x;\theta) \\ & - \log \hat{p}(y) - \log p(x|y;\phi) ||\,, \end{aligned} \tag{4}$$

here we use the estimated marginal distribution $\hat{p}(x)$ and $\hat{p}(y)$ instead.[2]

4. Dually-Syntactic Structure Matching

In this section we focus on the text↔text scenario, and present a dually-syntactic structure matching method for dual learning. We first demonstrate the detailed method at §4.1, and then present the experiments at §4.2.

4.1. Method

Dually-Syntactic Structure Encoding The input for both the primal and dual task is sentential words $\{w_1, \cdots, w_n\}$. Meanwhile we have its syntactic constituency parse $\mathcal{T} = \{T_k\}_{k=1}^K$, where T_i is an intermediate constituency phrase or terminal word, and K denotes the total node number. Constituency syntax describes the way the words compose into phrases and the tree, and such phrasal composition characteristic well fits our need for structure matching. At the encoding phase, the input words are mapped into contextual representations $\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n\}$ via a certain text encoder, e.g., BiLSTM, BERT. Then we use a tree model to encode the word representations into structure representations

[2]Details are elaborated at Appendix A.2.

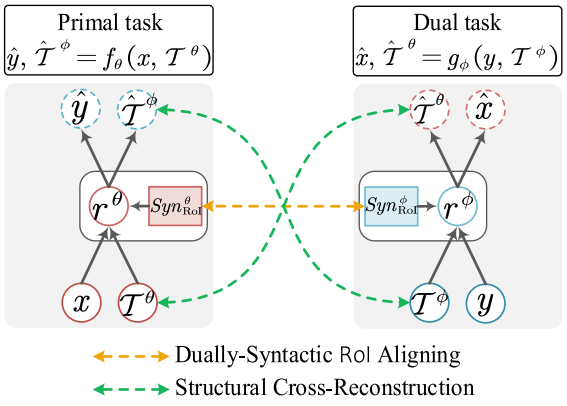


Figure 2. Symmetrically syntactic structure matching for dual learning.

$\boldsymbol{R} = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_K\}$, according to the constituent structure $\mathcal{T}$. Here we take the N-*ary* TreeLSTM (Tai et al., 2015) as the structure encoder. Without losing generality, we denote the node representation of constituency structure for primal task as $\boldsymbol{R}^\theta$, for dual task as $\boldsymbol{R}^\phi$.

Syntactic RoI Alignment The core idea is to build the fine-grained structure correspondences between primal and dual tasks, pushing those pairs that serve the similar role in the context to be closer, i.e., $p(T_i|\mathcal{T}^\theta) \approx p(T_j|\mathcal{T}^\phi)$. Specifically,

$$p(T_i|\mathcal{T}) = \text{Sigmoid}(\text{FFNs}(\text{Att}(T_i|\mathcal{T})))\,,$$

where $\text{Att}(\cdot)$ is an attention operation:

$$\text{Att}(T_i|\mathcal{T}) = \sum_{j=1, j\neq i}^{k=K} \beta_j \boldsymbol{r}_j\,,$$
$$\beta_j = \text{Softmax}(\boldsymbol{V}^T[\boldsymbol{r}_i; \boldsymbol{r}_j])\,,$$

here $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ are the representation of RoI T_i and T_j in $\mathcal{T}$.

Note that only a subset of the structure plays the pivotal role as rationale for such alignment, i.e., RoI. Technically, we first compute alignment scores between all pairs of constituents from two sides:

$$s_{i,j} = \frac{(\boldsymbol{r}_i^\theta)^T \cdot \boldsymbol{r}_j^\phi}{||\boldsymbol{r}_i^\theta||\,||\boldsymbol{r}_j^\phi||}\,. \tag{5}$$

With this we obtain a bipartite alignment in between. Also via a threshold ω we filter out those non-salient alignments with lower confidence, i.e., $p(T_i|\mathcal{T}) < \omega$, and obtain the candidate RoI pairs, i.e. Syn^θ_{RoI} and Syn^ϕ_{RoI}. A ranking loss is then used to pull closer those RoI pairs with higher similarities:

$$\mathcal{L}_M = \begin{cases} |s_{i,j}|\,, & s_{i,j} > \sigma \\ max(0, M - |s_{i,k}|)\,, & s_{i,j} \leq \sigma \end{cases} \tag{6}$$

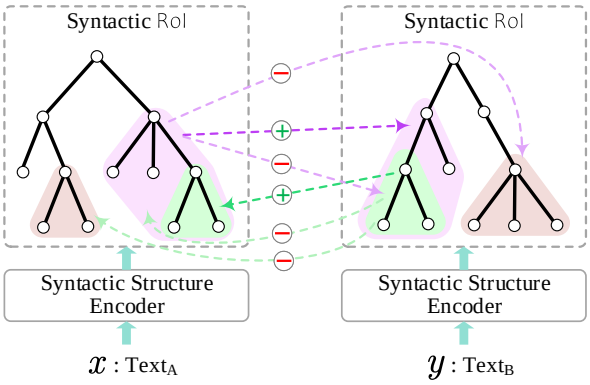


Figure 3. Dually-syntactic RoI alignment.

where M is a margin value, $\sigma \in [0, 1]$ is a self-adaptive threshold that is trainable during learning.

Contrastive Region Repelling Taking one step further, we make use of the negative samples; we hope the regions T_i in $\mathcal{T}^\theta$ that gives lower similarities to the one T_k in $\mathcal{T}^\phi$ to repel each other. Inspired by recent success of contrastive representation learning (Logeswaran & Lee, 2018; Giorgi et al., 2021), we replace the ranking loss with:

$$\mathcal{L}_M = -\sum_{i\in\mathcal{T}^\theta,\, j^*\in\mathcal{T}^\phi} \log \frac{\exp(s_{i,j^*}/\tau)}{\mathcal{Z}}\,, \tag{7}$$

$$\mathcal{Z} = \sum_{i\in\mathcal{T}^\theta,\, k\in\mathcal{T}^\phi,\, k\neq j^*} \exp(s_{i,k}/\tau)\,, \tag{8}$$

where $\tau > 0$ is a annealing factor. j^* means a positive pair with i, i.e., $s_{i,j^*} > \sigma$. Figure 3 show the technical illustration of the dually-syntactic RoI alignment for text↔text dual learning.

Structural Cross-Reconstruction On the other hand, during the text generation of $\hat{y}$ we make the model meanwhile to reproduce the corresponding syntax tree structure $\hat{\mathcal{T}}^\theta$. The syntax structure of the inut text from the opposite side (i.e, $\mathcal{T}^\theta$) can serve as a supervised signal. The benefits of such structural cross-reconstruction are multiple: making the structural awareness in the dual modeling more sufficient, providing additional syntactic contriant for the procedure, and also ensuring a global view during the generation. We adopt the representative graph-based method for constituency parsing (Stern et al., 2017). The process is to measure the score of each span (i, j) to be a valid constituency phrase as well as the constituent label, based on the decoder representations $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n\}$.[3] We can summarize the learning objectives for structure reconstructions in primal and dual tasks:

$$\mathcal{L}_R = \mathcal{L}_R^\theta + \mathcal{L}_R^\phi\,. \tag{9}$$

Overall Optimization Figure 2 illustrates the overall input and output of the dual systems with dually-syntactic structure matching. Putting them all together, the joint learning target becomes:

$$\mathcal{L}(\theta, \phi) = \mathcal{L}_C + \lambda_1\mathcal{L}_D + \lambda_2\mathcal{L}_M + \lambda_3\mathcal{L}_R\,, \tag{10}$$

where λ_* refers to a specific coupling co-efficiency. Note that all the stories of the structural matching happens at training stage. During inference, the primal and dual tasks make their own prediction without attendance of the opposite task.

4.2. Exp-I: Text↔Text Applications

We examine the usefulness of the dually-syntactic structure matching for text↔text dual learning.

Setups We use StanfordNLP (Qi et al., 2018) to tokenize and lemmatize the texts, and parse the syntactic constituency trees. We consider two typical text-text generation tasks: NMT and paraphrase generation. For NMT, we use the WMT14 EN-DE and EN-FR data, and take the ParaNMT and QUORA datasets for paraphrase generation. We make comparisons among four methods:

- **M1**: ordinary singleton task scheme.
- **M2**: ordinary singleton task scheme encoding external constituency syntax feature.
- **M3**: vanilla dual learning scheme.
- **M4**: dual learning scheme with our proposed syntactic structure matching.

We also ablate our M4 by 1) only encoding syntax feature without matching (OnlySyn); 2) removing the RoI alignment (-SAln); 3) without the syntax reconstruction (-SyRec); 4) changing the aligning algorithm (ranking or contrastive learning) for RoI alignment.

We take the Transformer-based (Vaswani et al., 2017) generation architecture. For NMT, we also test with the seq2seq-based architecture. For paraphrase generation we additionally use the BART PLM representations (Lewis et al., 2020). Also we compare with several existing strong-performing works for these tasks. For NMT, we have **B1** (Xia et al., 2018), **B2** (Chen et al., 2019a) and **B3** (Wang et al., 2021). For paraphrase generation we include **B1** (Iyyer et al., 2018), **B2** (Gupta et al., 2018), **B3** (Chen et al., 2019b) and **B4** (Kumar et al., 2020). We additionally reimplement these baselines to obtain the reversed task results. For all experiments, we report the average scores along with the unbiased standard deviations on five runs with different random seeds.

[3] Appendix §A.5 shows details of constituency parsing.

		WMT14 (EN-DE)				WMT14 (EN-FR)			
		EN→DE		EN←DE		EN→FR		EN←FR	
• *Baseline*	B1	28.04	/	30.91	/	39.44	/	35.32	/
	B2	28.22	/	30.72	/	39.68	/	35.90	/
	B3	28.57	/	31.00	/	39.80	/	35.85	/
• *Seq2seq-based*	M1	16.24	/	20.69	/	29.92	/	27.49	/
	M2	17.06	+0.82	21.62	+0.93	31.15	+1.23	28.82	+1.33
	M3	16.81	/	20.81	/	31.99	/	28.35	/
	M4	**19.52**	**+2.71**	**23.24**	**+2.43**	**35.85**	**+3.86**	**31.27**	**+1.92**
• *Transformer-based*	M1	25.24	/	28.42	/	37.21	/	32.08	/
	M2	27.07	+1.83	29.84	+1.42	38.73	+1.52	33.95	+1.87
	M3	26.46	/	29.17	/	38.10	/	32.52	/
	M4(RANK)	29.71	+3.25	33.40	+ 4.23	42.28	+4.18	37.09	+4.57
	M4(CL)	**30.03**	**+3.57**	**33.96**	**+4.79**	**42.82**	**+4.72**	**37.76**	**+5.24**
	ONLYSYN	27.90	+1.44	30.81	+1.64	39.03	+0.93	34.60	+2.08
	-SALN	28.23	+1.77	31.15	+1.98	39.55	+1.45	35.07	+2.55
	-SYREC	29.56	+3.10	32.68	+3.51	41.17	+3.07	36.34	+3.82

Table 1. Results (BLEU scores) on NMT. Two colors indicate the coupled tasks, respectively. Color depth highlights the significance of the result improvements. '+' means the improvement over the counterpart without using structure knowledge (e.g., M2-M1, M4-M3).

		ParaNMT						QUORA					
		B	R-1	R-2	R-L	B	R-1	B	R-1	R-2	R-L	B	R-1
• *Baseline*	B1	20.4	50.3	25.2	51.6	21.8	46.4	19.5	40.6	22.5	44.6	17.8	44.1
	B2	20.8	49.6	28.4	48.6	19.0	45.0	22.3	56.4	26.2	52.3	21.0	52.8
	B3	23.6	54.8	32.0	58.3	25.4	48.7	30.4	62.6	42.7	65.4	28.1	60.5
	B4	27.5	60.6	36.9	54.5	27.2	53.2	35.8	68.1	45.7	70.2	35.6	65.7
• *Transformer-based*	M1	24.6	50.3	30.7	45.8	25.4	51.7	29.7	58.5	37.5	59.6	28.0	60.5
	M2	27.2	56.4	34.4	50.6	26.1	53.6	33.4	63.4	41.8	63.4	34.8	65.8
	M3	26.2	57.1	33.0	53.5	27.8	55.9	32.0	65.7	40.0	66.4	34.0	64.3
	M4(RANK)	30.1	61.8	38.9	59.8	30.2	62.5	37.3	70.4	47.2	72.4	37.4	71.2
	M4(CL)	**30.5**	**62.4**	**39.4**	**60.4**	**30.6**	**62.7**	**37.5**	**70.5**	**47.6**	**72.5**	**37.5**	**71.5**
	ONLYSYN	27.7	58.9	34.9	54.7	28.0	56.2	33.7	66.4	42.0	67.1	35.0	65.8
	-SALN	28.0	59.6	35.8	56.0	28.6	57.3	34.6	67.6	43.2	68.9	35.8	67.4
	-SYREC	29.7	60.2	37.8	58.3	29.7	61.0	36.1	68.9	45.0	71.4	36.5	69.3
	M3+BART	33.8	65.7	41.8	62.8	32.7	64.0	41.5	73.3	49.4	74.2	42.0	71.5
	M4+BART	**36.7**	**66.2**	**43.6**	**64.0**	**34.8**	**64.6**	**43.0**	**74.8**	**52.8**	**76.8**	**43.5**	**72.8**

Table 2. Results on paraphrase generation (SRC→TGT, SRC←TGT). B: BLEU, R-X: ROUGE-X.

Results From the results and trends shown in Table 1 and 2 we have the following observations.

First of all, by comparing M2 to M1 and M4 to M3 we learn that the integration of syntactic structure results in better performances, either for the singleton or dual learning. Then, by comparing M3 to M1, it is clear that the dual learning technique improves the task performances consistently. Such improvements can been witnessed in both the primal and dual tasks. Third, when performing the proposed RoI matching (M4 vs. ONLYSYN), the vanilla dual learning scheme receives very significant enhancements over four datasets on all metrics, more than any other factors. This proves the efficacy of our structural matching proposal for text↔text dual learning.

Further, let's step into the RoI matching itself. Comparing the RoI alignment and syntactic structure reconstruction (M4-SALN vs. M4-SYREC), the former plays the predominant influences to the entire method. Also, the contrastive learning can bring better effectiveness than the ranking loss (M4(CL) vs. M4(RANK)), when performing the RoI alignment. This demonstrates the necessity to make use of the negative samples for the alignment.

Besides, we see that our model (M4) beat all comparing methods on all tasks and data, including the best-performing baselines. Also we can notice that, using the pre-trained con-

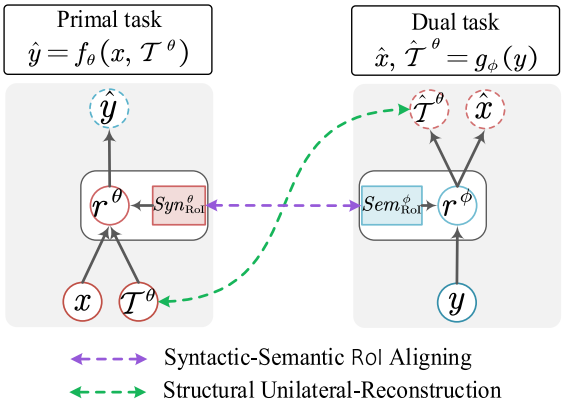


Figure 4. Syntactic-semantic structure matching.

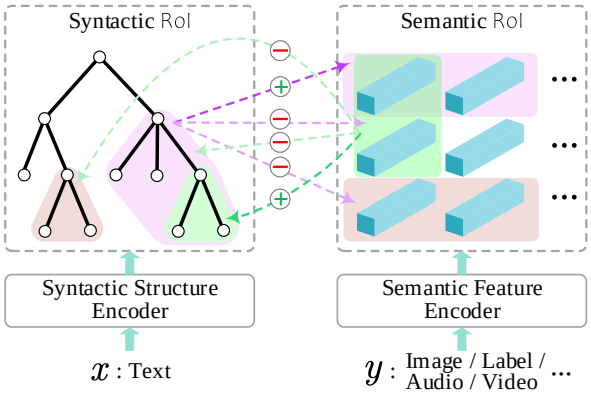


Figure 5. Syntactic-semantic RoI alignment via contrastive representation learning.

textualized word representations (i.e., BART), the improvements by our structure matching strategy can be slightly limited, even though our method (M4) still keeps the best. The possible reason can be that, BART already brings rich external features for enhancing the text understanding, in which the assistance for achiving better representation learning by our method of structural matching could consequently be weakened somewhat.

5. Syntactic-Semantic Structure Matching

In fact, it can be a broader range of NLP scenarios with dual learning technique where the task pair often includes non-text modalities, such as labels, image or audio etc. This makes the structure matching idea for text↔non-text dual learning non-trivial. This section will natrually extends the above method of dually-syntactic structure matching to a method of syntactic-semantic structure matching.

5.1. Method

Since the task of non-text modality comes without explicit syntactic structure, our main idea is to take the semantic structure of non-text, and perform **syntactic-semantic RoI alignment** instead. Meanwhile, the syntactic structure reconstruction for the global-level benefit becomes **structural unilateral-reconstruction**. We show the schematic design in Figure 4.

Let's say in the dual system, the primal task coming with text input; the dual task has non-text as input. Extending the spirit in text↔text, here we on the one hand encode the external syntactic structure for the textual part, and yield structure representations $\boldsymbol{R}^\theta$. On the other hand, we employ a semantic feature encoder for the non-text part. For example, for the image input we employ a object detector to generate a set of object proposals $\mathcal{O} = \{O_g\}_{g=1}^G$ and the corresponding vectorial embedding $\boldsymbol{R}^\phi$.

Likewise, we first calculate the relatedness between each pair of syntactic region and semantic region. Instead of directly taking the Cosine similarity as in Eq. (5), following Wang et al. (2020) we use a non-linear transformation for the scoring, as it is naturally a gap between the meaning spaces of different modalities:

$$s_{i,j} = \boldsymbol{V}^T(\boldsymbol{W}^\theta \boldsymbol{r}_i^\theta + \boldsymbol{W}^\phi \boldsymbol{r}_j^\phi)\,. \tag{11}$$

Next, the automatic threshold ω filters out invalid aligments, and yields the condidate RoI pairs, i.e. Syn_{RoI}^θ and Sem_{RoI}^ϕ. Here we inherit the success of the foregoing contrastive learning (Eq. 7), and perform the syntactic-semantic RoI alignment, as illustrated in Figure 5.

Meanwhile, we perform structural unilateral-reconstruction, i.e., letting the dual task at the same time generate text $\hat{x}$ and $\hat{\mathcal{T}}^\theta$ from the guidance of primal task's input. The objective for the dual learning system with syntactic-semantic structure matching is aligned with the prior one:

$$\mathcal{L}(\theta, \phi) = \mathcal{L}_C + \lambda_1 \mathcal{L}_D + \lambda_2 \mathcal{L}_M + \lambda_3 \mathcal{L}_R\,. \tag{12}$$

5.2. Exp-II: Text↔Non-Text Applications

Here we present the evaluations of our method in this section for text↔non-text scenarios.

Setups We mainly consider two cases of text↔image and text↔label, which represent two common dual learning applications. For text↔image, we take the MSCOCO and Flickr30k datasets. For text↔label we use the Yelp2014 and IMDB datasets. The settings of the comparing systems are kept the same as in §4, including **M1**, **M2**, **M3** and **M4**. The text→image backbone architecture is ControlGAN (Li et al., 2019), and text←image is BUTD (Anderson et al., 2018). The text→label backbone is Transformer and text←label is the VAE model (John et al., 2019). For more experimental setups and implementations, please refer to Appendix §A.

	MsCoCo				Flickr30k			
	IS↑	FID↓	B-4	MTR	IS↑	FID↓	B-4	MTR
M1	25.6	28.3	32.5	22.8	6.8	36.8	17.6	15.5
M2	27.8	25.5	/	/	7.5	35.0	/	/
M3	28.4	24.8	36.1	25.1	7.3	34.2	20.1	17.2
M4	**30.7**	**20.6**	**40.0**	**29.6**	**8.0**	**30.9**	**22.6**	**19.5**
-SAln	29.0	21.5	37.3	28.3	7.4	33.0	21.3	17.9
-SyRec	29.8	21.3	39.2	29.0	7.7	31.8	21.9	18.6

Table 3. Results on text↔image experiment (TXT→IMG: text-to-image synthesis, TXT←IMG: image captioning). B-4: BLEU-4, MTR: METEOR.

	Yelp2014				IMDB			
	ACC	B-4	MTR	ACC	ACC	B-4	MTR	ACC
M1	60.6	17.8	33.0	53.8	50.6	17.6	36.9	43.6
M2	61.8	/	/	/	51.9	/	/	/
M3	62.0	19.4	36.4	56.6	53.8	18.3	41.4	47.3
M4	**63.8**	**21.8**	**40.8**	**62.4**	**55.6**	**20.2**	**47.1**	**50.9**
-SAln	63.2	19.9	37.0	57.2	54.2	18.9	44.6	48.4
-SyRec	62.9	20.4	38.5	61.8	55.0	19.5	46.0	49.3

Table 4. Results on Text↔Label experiment (TXT→LB: text classification, TXT←LB: conditioned text generation).

Results Table 3 and 4 present the results. We find that the overall trends are kept similar with the ones for the text↔text cases. The syntactic-semantic structure matching strategy brings significant improvements for the vanilla dual learning across total 4 datasets and 8 tasks consistently. This means that the success of our proposed method can be inherited to the dual learning scenarios more than purely texts. The ablation studies for the syntactic-semantic RoI alignment mechanism and structural unilateral-reconstruction also show the homologous trends as in the foregoing cases.

6. Analysis and Discussion

Previously via numerical evaluations we have verified the efficacy of the structure matching for dual learning. Here we further explore several pivotal questions to better understand its strengths.

Questions

★ First, how does structure matching strategy improve the dual learning?

★ Second, for the text generation what are improved when aligning the structures?

★ Third, can the success of the structure alignment be extented to fully non-text scenarios?

★ Fourth, what are the key factors to the structure matching for dual learning?

6.1. Evaluating Structure Matching

Following we examine the underlying machnism of the structure matching, under the scenario of text↔text and text↔image cases respectively.

► **Structure matching helps correctly retrieve and emphasize the key RoIs that are crucial to the task improvements.** To evaluate the unsupervised matching correctness of our method, we first construct labels for each data, where *the key correspondences of pairs between the coupled tasks* are explicitly annotated as 'gold supervision'.[4] In the contrast experiments, we modify the syntactic-enhanced dual learning models by emphysizing the syntactic encoding with the 'gold RoI'.[5] Structure alignment enhances the correspondence of the critical feature region (e.g., textual spans or image regions) between two dual processes. Intuitively, more precise of the RoI alignments, the higher the improvements for the dual systems (Xia et al., 2017; Wang et al., 2020). The results in Table 5 show that our method can automatically learn the key RoI precisely, i.e., with quite small gaps to the results that use the gold RoIs.

	WMT14 (EN-DE)		WMT14 (EN-FR)	
	EN→DE	EN←DE	EN→FR	EN←FR
+ Auto RoI	29.03	31.96	41.82	36.76
+ Gold RoI	**29.51**	**32.23**	**42.03**	**36.98**
Δ	-0.48	-0.27	-0.31	-0.22
	ParaNMT		QUORA	
	SRC→TGT	SRC←TGT	SRC→TGT	SRC←TGT
+ Auto RoI	31.53	30.60	38.66	37.58
+ Gold RoI	**31.86**	**30.85**	**39.02**	**38.11**
Δ	-0.33	-0.25	-0.36	-0.53

Table 5. Results (BLEU) of dual learning with automatically learned and gold RoI matching respectively.

Taking a step further, we test how exactly correct our method can align the key RoIs. We evaluate the RoI matching correctness of our method on dual learning, where the results for text↔text generation are shown in Figure 6, and the results for text↔image task are as in Table 6.[6] And we see that our method (StruMatchDL) achieves over 85% accuracies comparing with the gold RoI matching in text↔text. Without the RoI alignment (Eq. 7), i.e, with only the structure reconstruction, the matching effectiveness can be greatly weakened. Comparatively, the influences from the structure reconstruction are much milder. We also see from the text↔image case that our StruMatchDL unsupervisedly learns good text-visual alignments, which are slightly lower

[4] The annotation details are shown in Appendix §A.7.
[5] Appendix §A.8 gives full model descriptions.
[6] Appendix §A.9 details the matching evaluation setups.

than the supervised visual grounding system.

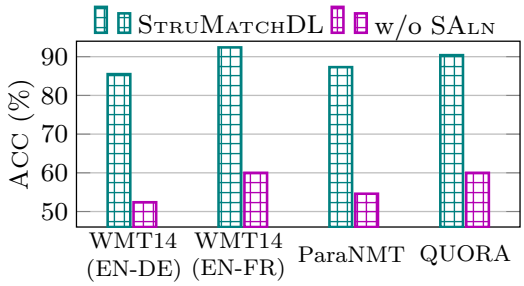
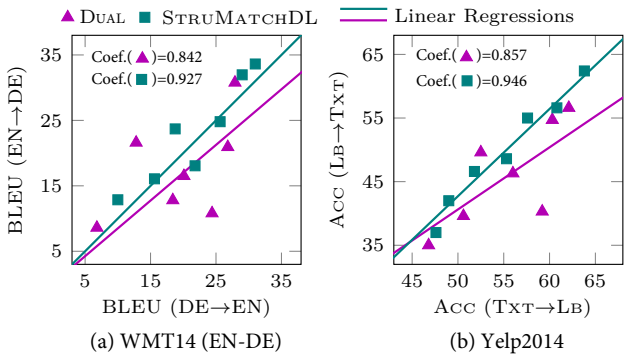


Figure 6. Measuring text↔text `RoI` alignment.

	ACC
MAF	61.4
StruMatchDL	**54.3** ± **0.3**
-SyRec	46.7 ± 0.5
-SAln	28.6 ± 0.8

Table 6. Visual grounding results on `Flickr30k` test set for verifying text↔image matching. MAF is a supervised visual grounding system (Wang et al., 2020).

► **Our method strengthens the duality between two dual tasks by correctly aligning the `RoIs`.** We further plot the performance correlation between the coupled tasks in Figure 7. The testing results are varied by using different proportion of training data. By comparing the linear regressions of the trends respectively, we see that the dual learning systems with structure matching show higher task correlations.



(a) WMT14 (EN-DE) (b) Yelp2014

Figure 7. Performance correlation between two coupled tasks. 'Coef.' indicates Pearson correlation coefficient.

6.2. Evaluating Generated Text

► **With syntactic structure co-echoing between the text↔text dual learning, the generated sentences are more diversified and grammarly correct.** As the major focus of this work, the text generation are significantly enhanced by the structure matching algorithm in dual learning

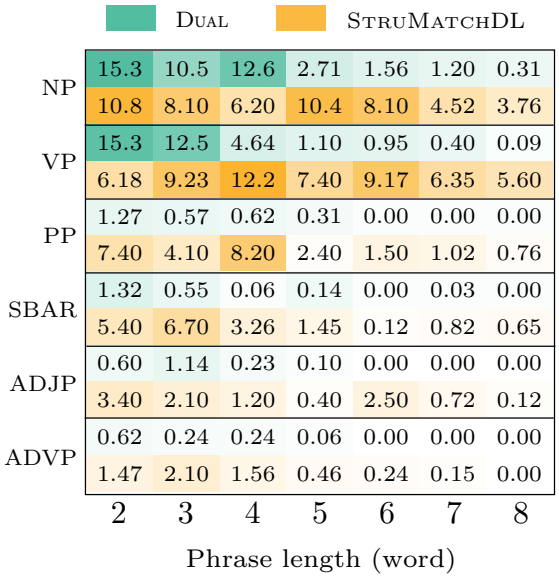


Figure 8. Distribution (frequency, %) over different constituency length of phrases in the generated sentences.

	ParaNMT			MsCoCo		
	Gram.	Corr.	Cont.	Gram.	Corr.	Cont.
Human	4.86	4.92	3.78	4.82	4.15	4.37
Baseline	1.58	2.20	1.04	0.78	1.23	0.98
Dual	2.24	2.55	1.46	1.80	2.38	1.25
StruMatchDL	**3.78***	**3.67***	2.51	**3.46***	**3.27***	2.74
-SyRec	2.89	3.21	**2.90***	2.75	2.89	**2.96***

Table 7. Human evaluation results. Grammaticality (Gram.), correctness (Corr.), and content richness (Cont.) are rated on Likert 5-scale. $*$ indicates significantly better over the variant ($p<0.03$).

system. Here we observe the details, figuring out what are really improved. First, in Figure 8 we plot the phrase type distribution over different constituent length in the generated texts on the two paragraph generation datasets. Comparing with vanilla dual learning that generates short phrases, our model helps yield a more even and smooth distribution of the phrase types and comparatively longer phrases.

Further, we ask five proficient English speakers to assess the quality of generated texts, where the results are shown in Table 7. We see that our method helps produce more correct generations both in content and grammarly, compared to the vanilla dual learning. Also, the syntactic structure ensures better content planning and better diversification, which are in line with relevant findings (Kumar et al., 2020; Bugliarello & Elliott, 2021). Interestingly, we find that even the syntactic structure reconstruction contributes to the overall better results, but in the cost of hurting some diversifications to certain extents.

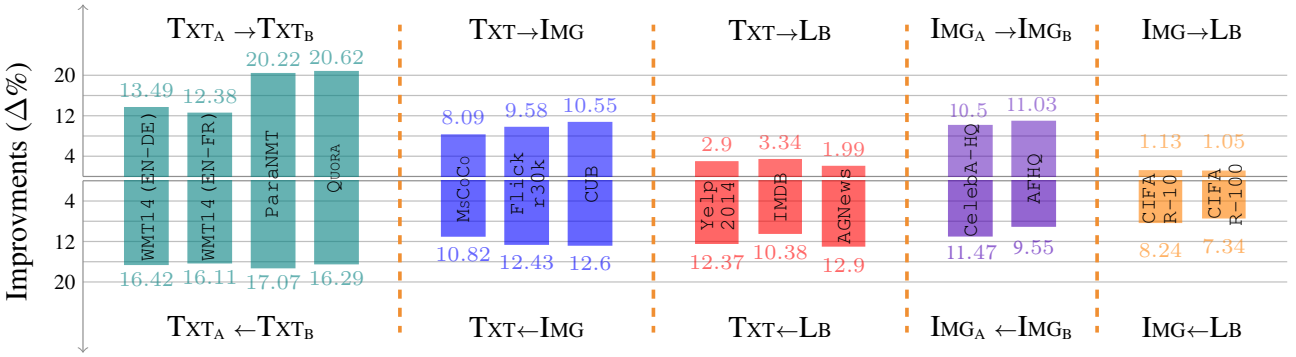


Figure 9. Relative task performance growth rates (Δ%) after taking the structure matching for dual learning.

6.3. Exploring Extendibility

► **Non-text↔non-text dual learning can also benefit from structure matching.** The prior experimental results raise further questions: what if both two sides of tasks comes without explicit syntax structure, i.e., taking the semantic-semantic structure alignment. Here we consider performing the evaluation on two representative tasks: image↔image (image-image translation) and image↔label (image classification vs. conditioned image generation). We can only take the semantic-semantic `RoI` alignment. We report the experiments at Appendix §B.1, where the results over four datasets clearly show that the improvements can be still retained by our method. For example, the rates of performance rises for image-image translation are over 9.55% as in Figure 9. However, without the explicit structure reconstruction at decoding side, the performance raises are not as significant as that with the reconstruction.

6.4. Insights into Key Influencers

► **The dual tasks with richer structural information for the alignments will lead to better improvements.** We further dig into the task improvements of all the dual learning tasks by our structure matching strategy. In Figure 9 we present the result growth rate on each task.[7] The comparisons evidently show that those alignments in the dual framework that come with richer structural information can provide higher task enhancements. For example, texts and images carry ampler (constituency or visual regions) structure than the labels, and thus the text↔text, text↔image and image↔image receive bigger raises. Also it is interesting to see that, when one task (say $\mathcal{A}$) comes with richer structure information than that in the opposite one (say $\mathcal{B}$), $\mathcal{B}$ can benefit more from $\mathcal{A}$. The text↔label and image↔label prove such case, where the label→* tasks gain much more improvements in the dual learning.

[7] We further add experiments on two text↔image and two text↔label datasets, cf. Appendix §B.2.

7. Conclusion and Future Work

In this work we investigate a structure matching mechanism for enhancing the duality in dual learning systems. We propose aligning the syntactic region of interests (`RoIs`) between two input sentences of two coupled tasks in the text↔text dual systems. The syntactic structure reconstruction at decoding phase is also performed to enhance the structural awareness. We then extend the structure matching to the text↔non-text scenarios with the syntactic-semantic structure alignment. We demonstrate the efficacy of the structure matching algorithm on a wide range of dual learning applications and datasets. We prove that our proposal helps correctly retrieve and reinforce the key structure regions that are critical to the task improvements. Finally, we reveal the great potential of the structure alignment for many other non-text↔non-text dual learning scenarios.

This work may limit within the scope of supervised dual learning. Meanwhile, we make use of the external parse trees as structural supervisions being encoded by a tree encoder for the structure alignment. This pipeline process may potentially introduce task-irrelevant noises. As a future work, we intend to automatically & unsupervised induce structural representations and simultaneously match the structures for both the supervised unsupervised setups of dual learning.

Acknowledgements

We thank the anonymous reviewers for their interesting suggestions. This research is supported by the Sea-NExT Joint Lab, the Key Project of State Language Commission of China (No. ZDI135-112), the Science of Technology Project of GuangZhou (No. 20210202607), and the National Natural Science Foundation of China (No. 62176180).

References

Aarts, F. and Aarts, J. M. *English syntactic structures: functions and categories in sentence analysis*, volume 1. Pergamon, 1982.

Akoury, N., Krishna, K., and Iyyer, M. Syntactically supervised transformers for faster neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1269–1281, 2019.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

Aneja, J., Deshpande, A., and Schwing, A. G. Convolutional image captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5561–5570, 2018.

Asghar, N. Yelp dataset challenge: Review rating prediction. *CoRR*, abs/1605.05362, 2016.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, 2015.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pp. 12–58, 2014.

Bugliarello, E. and Elliott, D. The role of syntactic planning in compositional image captioning. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pp. 593–607, 2021.

Cao, R., Zhu, S., Liu, C., Li, J., and Yu, K. Semantic parsing with dual learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 51–64, 2019.

Cao, S. and Wang, L. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pp. 6424–6439, 2021.

Chen, K., Wang, R., Utiyama, M., and Sumita, E. Neural machine translation with reordering embeddings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1787–1799, 2019a.

Chen, M., Tang, Q., Wiseman, S., and Gimpel, K. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 5972–5984, 2019b.

Choi, Y., Uh, Y., Yoo, J., and Ha, J. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8185–8194, 2020.

Corso, G. M. D., Gulli, A., and Romani, F. Ranking a stream of news. In *Proceedings of the international conference on World Wide Web*, pp. 97–106, 2005.

Dyer, C., Chahuneau, V., and Smith, N. A. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–648, 2013.

Eriguchi, A., Hashimoto, K., and Tsuruoka, Y. Tree-to-sequence attentional neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 823–833, 2016.

Fei, H., Ren, Y., and Ji, D. Improving text understanding via deep syntax-semantics communication. In *Proceedings of findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 84–93, 2020a.

Fei, H., Zhang, M., and Ji, D. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7014–7026, 2020b.

Fei, H., Li, F., Li, B., and Ji, D. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12794–12802, 2021a.

Fei, H., Wu, S., Ren, Y., Li, F., and Ji, D. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP*, pp. 549–559, 2021b.

Gardner, M., Dasigi, P., Iyer, S., Suhr, A., and Zettlemoyer, L. Neural semantic parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 17–18, 2018.

Garmash, E. and Monz, C. Bilingual structured language models for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2398–2408, 2015.

Giorgi, J., Nitski, O., Wang, B., and Bader, G. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pp. 879–895, 2021.

Goyal, T. and Durrett, G. Neural syntactic preordering for controlled paraphrase generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 238–252, 2020.

Gupta, A., Agarwal, A., Singh, P., and Rai, P. A deep generative framework for paraphrase generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5149–5156, 2018.

Hays, D. G. Dependency theory: A formalism and some observations. *Language*, 40(4):511–525, 1964.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., and Ma, W. Dual learning for machine translation. In *Proceedings of Annual Conference on Neural Information Processing Systems*, pp. 820–828, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016b.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., and Xing, E. P. Toward controlled generation of text. In *Proceedings of the International Conference on Machine Learning*, pp. 1587–1596, 2017.

Iyyer, M., Wieting, J., Gimpel, K., and Zettlemoyer, L. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1875–1885, 2018.

John, V., Mou, L., Bahuleyan, H., and Vechtomova, O. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 424–434, 2019.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations*, 2018.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kumar, A., Ahuja, K., Vadapalli, R., and Talukdar, P. P. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:330–345, 2020.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Li, B., Qi, X., Lukasiewicz, T., and Torr, P. H. S. Controllable text-to-image generation. In *Proceedings of Neural Information Processing Systems*, pp. 2063–2073, 2019.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In *Proceedings of Computer Vision of European Conference*, pp. 740–755, 2014.

Logeswaran, L. and Lee, H. An efficient framework for learning sentence representations. In *Proceedings of the International Conference on Learning Representations*, 2018.

Ma, S., Sun, X., Li, W., Li, S., Li, W., and Ren, X. Query and output: Generating words by querying distributed word representations for paraphrase generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 196–206, 2018.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, 2011.

Mao, Q., Lee, H., Tseng, H., Ma, S., and Yang, M. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1429–1437, 2019.

Marcheggiani, D. and Titov, I. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1506–1515, 2017.

Miao, N., Zhou, H., Mou, L., Yan, R., and Li, L. CGMH: constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6834–6842, 2019.

Miyato, T. and Koyama, M. cgans with projection discriminator. In *Proceedings of the International Conference on Learning Representations*, 2018.

Nguyen, T. H. and Shirai, K. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2509–2514, 2015.

Peng, G. and Wang, S. Dual semi-supervised learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8827–8834, 2019.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

Ponti, E. M., Reichart, R., Korhonen, A., and Vulić, I. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1531–1542, 2018.

Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. Universal Dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 160–170, 2018.

Ren, S., He, K., Girshick, R. B., and Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of the Neural Information Processing Systems*, pp. 91–99, 2015.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *Proceedings of the International Conference on Learning Representations*, 2017.

Shen, L. and Feng, Y. CDL: Curriculum dual learning for emotion-controllable response generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 556–566, 2020.

Shi, H., Mao, J., Gimpel, K., and Livescu, K. Visually grounded neural syntax acquisition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1842–1861, 2019.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

Stahlberg, F., Hasler, E., Waite, A., and Byrne, B. Syntactically guided neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 299–305, 2016.

Stern, M., Andreas, J., and Klein, D. A minimal span-based neural constituency parser. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 818–827, 2017.

Su, S.-Y., Huang, C.-W., and Chen, Y.-N. Dual supervised learning for natural language understanding and generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 5472–5477, 2019.

Sun, K., Zhang, R., Mensah, S., Mao, Y., and Liu, X. Aspect-level sentiment analysis via convolution over dependency tree. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 5679–5688, 2019.

Sun, Y., Tang, D., Duan, N., Qin, T., Liu, S., Yan, Z., Zhou, M., Lv, Y., Yin, W., Feng, X., Qin, B., and Liu, T. Joint learning of question answering and question generation. *IEEE Transactions on Knowledge and Data Engineering*, 32(5):971–982, 2020.

Tai, K. S., Socher, R., and Manning, C. D. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pp. 1556–1566, 2015.

Tjandra, A., Sakti, S., and Nakamura, S. Listening while speaking: Speech chain by deep learning. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 301–308, 2017.

van den Oord, A., Kalchbrenner, N., Espeholt, L., Kavukcuoglu, K., Vinyals, O., and Graves, A. Conditional image generation with pixelcnn decoders. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pp. 4790–4798, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pp. 5998–6008, 2017.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wang, F., Yan, J., Meng, F., and Zhou, J. Selective knowledge distillation for neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pp. 6456–6466, 2021.

Wang, K. and Wan, X. Automatic generation of sentimental texts via mixture adversarial networks. *Artifial Intelligence*, 275:540–558, 2019.

Wang, Q., Tan, H., Shen, S., Mahoney, M., and Yao, Z. MAF: Multimodal alignment framework for weakly-supervised phrase grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2030–2038, 2020.

Wieting, J. and Gimpel, K. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 451–462, 2018.

Wong, Y. W. and Mooney, R. Generation by inverting a semantic parser that uses statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 172–179, 2007.

Wu, S., Fei, H., Ren, Y., Ji, D., and Li, J. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 3957–3963, 2021.

Xia, Y., Qin, T., Chen, W., Bian, J., Yu, N., and Liu, T. Dual supervised learning. In *Proceedings of the International Conference on Machine Learning*, pp. 3789–3798, 2017.

Xia, Y., Tan, X., Tian, F., Qin, T., Yu, N., and Liu, T. Model-level dual learning. In *Proceedings of the International Conference on Machine Learning*, pp. 5379–5388, 2018.

Xu, X., Song, J., Lu, H., He, L., Yang, Y., and Shen, F. Dual learning for visual question generation. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2018.

Ye, H., Li, W., and Wang, L. Jointly learning semantic parser and natural language generator via dual information maximization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 2090–2101, 2019.

Zhang, S. and Bansal, M. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pp. 2495–2509, 2019.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations*, 2020.

Zhang, Y., Marshall, I., and Wallace, B. C. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 795–804, 2016.

Zhang, Y., Qi, P., and Manning, C. D. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2205–2215, 2018.

A. Details on Technical and Experimental Setups

A.1. Task Summary

Dual learning has a wide scope of applications between different modalities. In Table 8 we summarize some tasks and duality schemes that are used in this work.

Duality Scheme	Direction	Representative Application(s)
Text↔Text	$\longrightarrow$ or $\longleftarrow$	Neural Machine Translation, Paraphrase Generation
Text↔Image	$\longrightarrow$	Text-to-Image Synthesis
	$\longleftarrow$	Image Captioning
Text↔Label	$\longrightarrow$	Text Classification
	$\longleftarrow$	Conditioned Text Generation
Image↔Label	$\longrightarrow$	Image Classification
	$\longleftarrow$	Conditioned Image Generation
Image↔Image	$\longrightarrow$ or $\longleftarrow$	Image Translation

Table 8. Task summary in the dual viewpoint.

A.2. Marginal Distribution Estimation

A dual learning system seeks calculating the joint distribution as in Eq. (3). In the prime task side and the dual task side, the marginal distributions $p(x)$ and $p(y)$ are both required, but actually cannot be observed directly. Instead, we estimate these marginal distribution $p(x)$ of x (vice versa for $p(y)$) with a surrogate distribution $\hat{p}(x)$, by observing the target in the scope of the whole data.

- For the target of textual sentences, we use a Transformer-based state-of-the-art language model that is trained over the specific data to calculate the $\hat{p}(x)$ (Xia et al., 2017; Su et al., 2019).
- For the target of images, we follow Xia et al. (2017) and define the image distribution as $\hat{p}(x) = \prod_{i=1}^{m} px_i|x_{<i}$. We serialize the image pixels as x_i. We use PixelCNN++ (Salimans et al., 2017) to model this distribution.
- For the target of discrete labels, we simply use the uniform distribution of each class as $\hat{p}(x)$.

A.3. Relative Task Performance Improvement

In §6.3 we show the task improvements growth rates (Δ) over each prime-dual task pair when taking the structure matching strategy proposed in this paper. Specifically, the relative improvement is calculated as follow:

$$\Delta = \frac{||M_1 - M_2||}{M_1},$$

where M_1 refers to the performance (in one specific metric) of the vanilla dual learning system. M_2 denotes the performance of the dual learning system enhanced by structure matching. The primary metric (M) of each task that is used to calculate Δ is listed in Table 9.

Task	Direction	Metric
Text↔Text	$\longrightarrow$ or $\longleftarrow$	BLEU
Text↔Image	$\longrightarrow$	IS
	$\longleftarrow$	BLEU-4
Text↔Label	$\longrightarrow$	Accuracy
	$\longleftarrow$	Accuracy
Image↔Label	$\longrightarrow$	Accuracy
	$\longleftarrow$	IS
Image↔Image	$\longrightarrow$ or $\longleftarrow$	FID

Table 9. The primary metric of each task for calculating the Δ.

A.4. Implementation Detail of Model Architecture

There are some commonly used experimental setups. All the four comparing systems, i.e., **M1**, **M2**, **M3** and **M4**, use the same baseline architecture for the coupled dual tasks, for fairness. Also for **M2** and **M4** in text↔* scenarios, the syntactic structure encoder is kept the same, i.e., N-*ary* TreeLSTM (Tai et al., 2015). For all the experiments we take a two-layer TreeLSTM. The flow in TreeLSTM is made bidirectional, i.e., bottom-up and top-down, for a full information interaction. Also the TreeLSTM encodes the constituency label features.

Technically, for each node k in the tree, we denote the hidden state and memory cell of its v-th ($v \in [1, M]$) branching child as $\boldsymbol{r}_{kv}^{\uparrow}$ and $\boldsymbol{c}_{kv}$, and the embedding $\boldsymbol{h}_k^{\pi}$ of constituency label for node k. The bottom-up one computes the representation $\boldsymbol{r}_k^{\uparrow}$ from its children hierarchically:

$$\boldsymbol{i}_k = \sigma(\boldsymbol{W}^{(i)}[\boldsymbol{h}_k; \boldsymbol{h}_k^{\pi}] + \sum_{v=1}^{M} \boldsymbol{U}_v^{(i)} \boldsymbol{r}_{kv}^{\uparrow} + \boldsymbol{b}^{(i)}),$$

$$\boldsymbol{f}_{ku} = \sigma(\boldsymbol{W}^{(f)}[\boldsymbol{h}_k; \boldsymbol{h}_k^{\pi}] + \sum_{v=1}^{M} \boldsymbol{U}_{uv}^{(f)} \boldsymbol{r}_{kv}^{\uparrow} + \boldsymbol{b}^{(f)}),$$

$$\boldsymbol{o}_k = \sigma(\boldsymbol{W}^{(o)}[\boldsymbol{h}_k; \boldsymbol{h}_k^{\pi}] + \sum_{v=1}^{M} \boldsymbol{U}_v^{(o)} \boldsymbol{r}_{kv}^{\uparrow} + \boldsymbol{b}^{(o)}),$$

$$\boldsymbol{u}_k = \text{Tanh}(\boldsymbol{W}^{(u)}[\boldsymbol{h}_k; \boldsymbol{h}_k^{\pi}] + \sum_{v=1}^{M} \boldsymbol{U}_v^{(u)} \boldsymbol{r}_{kv}^{\uparrow} + \boldsymbol{b}^{(u)}),$$

$$\boldsymbol{c}_k = \boldsymbol{i}_k \odot \boldsymbol{u}_k + \sum_{v=1}^{M} \boldsymbol{f}_{kv} \odot \boldsymbol{c}_{kv},$$

$$\boldsymbol{r}_k^{\uparrow} = \boldsymbol{o}_k \odot \tanh(\boldsymbol{c}_k),$$

where [;] means the concatenation, $\boldsymbol{W}$, $\boldsymbol{U}$ and $\boldsymbol{b}$ are parameters. $\boldsymbol{h}_k$, $\boldsymbol{i}_k$, $\boldsymbol{o}_k$ and $\boldsymbol{f}_{ku}$ are the input token representation,

input gate, output gate and forget gate. Analogously, the top-down N-ary TreeLSTM calculates the representation $\boldsymbol{r}_k^{\downarrow}$ the same way. We concatenate the representations of two directions: $\boldsymbol{r}_k = [\boldsymbol{r}_k^{\uparrow}; \boldsymbol{r}_k^{\downarrow}]$.

Except the experiments of paraphrase generation where we use the pre-trained contextualized BART representation[8] for enhancements, in other experiments we only take the fasttext English word embedding[9], and most of the word embedding dimension is set as 300. For text generation tasks, the byte pair encoding (BPE) technique is used for subword segmentation, and the vocabulary size is set at 20∼40K depending on specific corpus.

Besides, for facilitating the grounding of image semantic feature, we use an external object detector, Faster R-CNN (Ren et al., 2015) to generate object proposals $\mathcal{O} = \{O_g\}_{g=1}^{G}$, and extract the visual features with `RoI`. Specifically, for all the following models for processing images, we replace the kernel feature encoder with R-CNN. For each visual `RoI` proposal, we use global average pooling to compute its conv-feature and embed it to a vector (He et al., 2017).

For all the dual learning system, we first pre-train the two standalone coupled models separately, and then perform the joint dual training, so as to avoid cold-start training that will cause unstable learning or even the failures of convergence. Once two models in the dual learning system have been well trained to reach the best validating performances, during inference these two models will then make predictions separately without relying on each other's representations for `RoI` matching.

1) Text↔Text Applications For the dual text-to-text translation/generation task, we adopt the standard Transformer (Vaswani et al., 2017) as baseline encoder or decoder. For the NMT, we take the 12-layer configuration, while for paraphrase generation we take the 6-layer of Transformer. Also we use the position embedding. Besides, for the NMT task, we meanwhile implement the sequence-to-sequence baseline, which is a standard attention-based encoder-decoder architecture (Bahdanau et al., 2015), with 3-layer BiLSTMs as encoder and 2-layer LSTM as decoder. We use beam search with a beam size 5 and length penalty 1.0, so as to yield 5-best generated texts.

2) Text↔Image Applications For the text-to-image synthesis, we adopt the standard ControlGAN (Li et al., 2019) as the backbone model, which is a state-of-the-art image generation system build upon the generative network. For the image caption task, we take the standard BUTD (Anderson et al., 2018), which is a strong and widely-employed RNN-based captioning system that implements the bottom-up and top-down attention mechanism. Also, the captioning employ the beam size of 5.

3) Text↔Label Applications The system for the text classification is the 3-layer Transformer model. We adopt the Adam optimizer with an initial learning rate of 1e-5 for training the classifier. For the conditioned text generation, we take the classical text VAE model (John et al., 2019) as the backbone architecture. The latent variable representation of text style (sentiment label) has 8 dimensions and the variable of content space has 128 dimensions, being same as in (John et al., 2019).

4) Image↔Label Applications For image classification, we take the standard ResNet-110 (He et al., 2016b), which is a high-performing visual handling system that uses the deeply-stacked convolution layers with residual connections. For the conditional image generation, we take the CGAN (Miyato & Koyama, 2018), where we use the same configurations as in the raw paper.

5) Image↔Image Applications We use the MSGAN (Mao et al., 2019) model as our image translation backbone architecture. We take the officially-released models as initiation for warm-start training, and all the settings for MSGAN are kept the same with the raw paper.

A.5. Constituency Parsing for Syntactic Structure Reconstruction

In §4.1 we propose performing syntactic structure reconstruction, i.e., generating $p(\hat{\mathcal{T}}^{\theta}|x)$, given the syntax tree structure from opposite side (i.e, $\mathcal{T}^{\theta}$) as supervision. We take the graph-based method for constituency parsing (Stern et al., 2017), measuring the score of each span (i, j) to be a valid constituency phrase as well as the constituent label. First, we based on the decoder representations $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n\}$ of each word token create all the text spans iteratively $e_{i,j}$, where i and j represent the start and end indexes of the span. Next we use two separate feedforward networks (FFNs) to obtain the scores of the span and its label:

$$s_{sp}(i, j) = \text{FFNs}^{span}(e_{i,j}) \,,$$
$$\boldsymbol{s}_{lb}(i, j) = \text{FFNs}^{label}(e_{i,j}) \,.$$

Here the label score $\boldsymbol{s}_{lb}(i, j)$ is a vector, with each dimension representing the score of the possiblity as the corresponding label l:

$$s_{lb}(i, j, l) = [\boldsymbol{s}_{lb}(i, j)]_l \,.$$

Then we use the CKY-based chart parsing to measure the

[8] BART base version, https://huggingface.co/facebook/bart-base

[9] https://fasttext.cc/

overall score of a tree:

$$s_{tree}(\mathcal{T}) = \sum_{(l,(i,j))\in\mathcal{T}} \left[s_{lb}(i,j,l) + s_{sp}(i,j)\right].$$

And the best tree with maximum score can be denoted as:

$$s_{best}(i,j) = \max_{l}[s_{lb}(i,j,l)] + \max_{k}[s_{split}(i,k,j) + s_{best}(i,k) + s_{best}(k,j)],$$

which is a dynmic programming problem. We take the max-margin as the training objective:

$$\mathcal{L}_R = \max(0, \Delta(\hat{\mathcal{T}}, \mathcal{T}^*) - s_{tree}(\mathcal{T}^*) + s_{tree}(\hat{\mathcal{T}})),$$

where $\hat{\mathcal{T}}$ is the gold tree supervision.

A.6. Data and Evaluation Description

Here we give a detailed description on the datasets and the evaluation settings we used. For all experiments, we report the average scores along with the unbiased standard deviations on five runs with different random seeds. For the improvements by our proposed model over baselines, we use the paired T-test to examine that the gains are statistically significant, with $p<0.03$ or $p<0.05$.

1) Text↔Text Applications

- `WMT14(EN-DE)` (Bojar et al., 2014) splits the total sentences into training (4.6M), developing (3K) and testing (2K).
- `WMT14(EN-FR)` (Bojar et al., 2014) splits the sentences into training (36 M), developing (2.6K) and testing (2.6K).
- `ParaNMT` (Wieting & Gimpel, 2018) contains 500K sentence-paraphrase pairs for training. And the rest 1,300 manually labeled sentence pair is further split into 800 test data and 500 dev data.
- `QUORA`[10] includes 146K parallel paraphrases, 3K and 30K paraphrase pairs are respectively used for validation and testing, following Miao et al. (2019).

For neural machine translation, we use the standard BLEU score as the evaluation metric. For paraphrase generation, we report the precesion-oriented BLEU, recall-oriented ROUGE-1, ROUGE-2 and ROUGE-L.

[10] https://www.kaggle.com/c/quora-question-pairs

2) Text↔Image Applications

- `MSCoCo` (Lin et al., 2014) contains 113,287 training images equipped with five sentences each, and 5,000 images for validation and test splits, respectively.
- `Flickr30k` (Plummer et al., 2015) data contains 224K phrases and 31K images in total, where each image will be associated with 5 captions and multiple localized bounding boxes. Following previous work, we use 30k images ramdomly selected from training set.
- `CUB` (Wah et al., 2011) dataset contains 200 classes and 11,788 bird images in total, with 10 visual description sentences for each image. There are 8,855 and 2,933 images for training and testing.

The evaluation metrics for text-to-image synthesis include the widely-employed *Inception Score* (IS) and *Fréchet Inception Distance* (FID). To measure the image captioning, we take the BLEU-4 and METEOR.

3) Text↔Label Applications

- `Yelp2014` (Asghar, 2016) is a 5-class sentiment classification data, where the splits of training, developing and testing is 184K, 23K and 23K.
- `IMDB` data (Maas et al., 2011) labels each text with a 10-scale sentiment label, including 67.9K training sentences, 8.5K developing sentences and 8.5K testing sentences.
- `AGNews` (Corso et al., 2005) is a topic classification data with 4 topics, containing 110K training sentences, 10K developing sentences and 7.6K testing sentences.

For the text classification, we employ the accuracy (ACC) metric. For the conditioned text generation, we follow previous work (Cao & Wang, 2021) and emply the BLEU-4 and METEOR. Besides, following Wang & Wan (2019), we take the generated texts as inputs, and use a well-finetuned BERT classifier to measure the texts against the gold label. Then we use the accuracy to measure such performance.

4) Image↔Label Applications

- `CIFAR-10` (Krizhevsky et al., 2009) consists of 60k 3x32x32 colour images in 10 classes, with 6K images per class. There are 50k training images and 10K test images.
- `CIFAR-100` (Krizhevsky et al., 2009) also contains 60K images as in CIFAR-10, but it has 100 classes

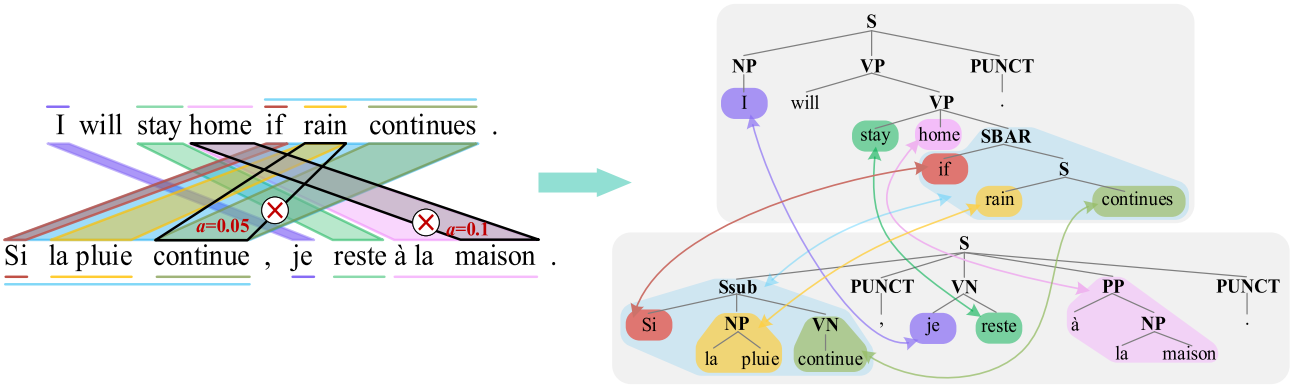


Figure 10. **Left**: chunk aligments between a pair of sentence (English ↔ French) for NMT. The regions with black box and red cross represent low-quality and invalid alignments, which will be rejected. **Right**: the alignments in the view of syntactic structure. For brevity, the POS tags and some redundant nodes are omitted.

containing 600 images each. There are 500 training images and 100 testing images per class. Each image comes with a fine label and a coarse label.

For the image classification, we take the accuracy, and for the conditioned image generation we use the IS and FID metrics.

5) Image↔Image Applications

- `CelebA-HQ` (Karras et al., 2018) contains 30k celebrity facial images, which are manually split into 17,943 female faces and 10,057 male faces for training, and the rest 2000 images are evenly divided as testing data, following Choi et al. (2020).
- `AFHQ` (Choi et al., 2020) includes 15k animal faces and is evenly distributed into three challenging domains, cat, dog and wildlife. Each domain uses 500 images for testing and the rest for training.

For the image-image translation, we take the FID evaluation metric.

A.7. Data Construction for `RoI` Alignment Evaluation

In §6.1 we examine the matching capability of the `RoI` between dual task pairs, by using the 'gold `RoI`' data. Here we illustrate how to construct such data for the Text↔Text and Text↔Image scenarios, respectively.

1) Dually-Syntactic Structure Pairing for Text↔Text Tasks

Step 1: Computing aligning score. For the sentence pair $(\{e_1, \cdots, e_n\} \leftrightarrow \{f_1 \cdots, f_m\})$ in NMT task, inspired by Fei et al. (2020b), we use the FastAlign tool (Dyer et al., 2013) to get the aligning probabilities $p(f_j|e_i)$ from the source word e_i to the target word f_j. Besides, we also use a Part-of-Speech (POS) tagger at target-side language to obtain the POS tag distribution $p(t_*|f_j)$ (t_* is an arbitrary POS tag). We then combine the two perspective into a total aligning score $a(e_i \leftrightarrow f_j) = p(f_j|e_i)p(t_*|f_j)$ for each word-pair to ensure a comprehensive alignment.

For the sentence pair $(\{e_1, \cdots, e_n\} \leftrightarrow \{f_1 \cdots, f_m\})$ in paraphrase generation task, we follow Zhang & Bansal (2019), and compute each alignment score $a(e_i \leftrightarrow f_j)$. The score is computed using a weighted mean of the contextual similarity between individual words. Specifically, the weights are the corpus-level inverse-document frequency (IDF) score d_e of the word e (same for target word f). We then use the contextual representation from BERT h_{e_i} to obtain the BERT-Score as in Zhang et al. (2020). Then, we compute the alignment between each word pair $e_i - f_j$:

$$a(e_i \leftrightarrow f_j) = \frac{1}{2}\Big(\frac{\sum_{e_i} d_{e_i} \max_{f_j} h_{e_i}^T h_{f_j}}{\sum_{e_i} d_{e_i}} + \frac{\sum_{f_j} d_{f_j} \max_{e_i} h_{f_j}^T h_{e_i}}{\sum_{f_j} d_{f_j}}\Big).$$

Step 2: Ensembling confidence of the `RoI`. Next we set a threshold value α to filter out those projections with low aligning confidence, i.e., $a(e_i \leftrightarrow f_j) < \alpha$. Note that the threshold varies depending on the specific task and data. So we obtain the partial mapping between each sentence pair. Then, we compose the words (with valid alignment) into phrases by joining the adjacent words.

Step 3: Aligning in the structure. We next project the flat alignment into a hierarchy of the constituency structure. We make pairing for the above valid paired phrases in the

A man with **pierced ears** is wearing **glasses** and **an orange hat**.

A man with **glasses** is wearing **a beer can crotched hat**.

A man with **gauges** and **glasses** is wearing **a Blitz hat**.

A man in **an orange hat** starring at something.

A man wears **an orange hat** and **glasses**.

Figure 11. **Left**: an iamge with regions of interest annotated in bounding boxes by different colores. **Right**: five captions describing the same image, where colored phrases link mentions of the same entities in image regions.

syntactic tree. In Figure 10 we demonstrate the overall alignment process with a NMT case.

2) Syntactic-Semantic Structure Pairing for Text↔Image Tasks

For the image captioning scenario, we use the `Flickr30k` data, where there are gold annotations for visual grounding. As shown in Figure 11, each key regions of image are grounded with the corresponding phrases in the caption. Note that in `Flickr30k` each image has an average of 5.26 captions sentences, and 7.7 critical regions of interest (grounding boxes). And we directly use such gold annotations for the evaluation of region-to-phrase correspondence.

A.8. Integration of Gold `RoI`

By using the above created 'gold' `RoI`[11], we can evaluate the efficacy of our structure alignment method, i.e., testing how accuracy it retrieves key `RoIs`, as shown in §6.1. The comparing setup is a model that takes such gold `RoIs` for explicit supervision, so as to learn a better feature representations for better task performances, and meanwhile achieving explicit alignments (i.e., via supervised learning).

The comparing dual learning model used here has a modified syntactic structure encoder, where specifically in the N-ary TreeLSTM those nodes representing key `RoIs` are highly weighted. For those nodes k as key `RoIs` in the constituency tree, the corresponding weights γ_k assigned are higher, and for those nodes k as not key `RoIs`, the weights are assigned much lower. Then, we only need to change the node representations:

$$\boldsymbol{r}_k := \gamma_k \cdot \boldsymbol{r}_k \,,$$

so as to control the influences of the nodes in the tree encoder. By highlighting the critical feature region (e.g., text spans or image regions) between two dual processes, the task performances as well as the correlations (or dualities) between the paired tasks can be enhanced, as in the Table 5.

[11] For the text↔text case, the generated labels of `RoI` are silver; for the text↔image case, the `RoI` are manually annotated gold labels.

A.9. Evaluation Detail of `RoI` Matching

In §6.1 we evaluate the `RoI` matching correctness of our method on dual text↔text generation as in Figure 6, and text↔image task as in Table 6. There we show the evaluation details.

1) on Text↔Text Tasks For the text↔text generation, we just take the automatically learned alignments from our model as our prediction of `RoIs`. We then make comparisons with the silver `RoIs` created at Appendix §A.7 on each datasets.

2) on Text↔Image Tasks For the text↔image task, we only need to evaluate the visual grounding performances on the `Flickr30k` test set. To show the strengths of our model, we additionally make comparisons with a supervised visual grounding system MAF (Wang et al., 2020).

B. Extended Experiments

B.1. Results on Non-Text↔Non-Text Applications

In §6.3 we explore the performances of the structure matching idea for the non-text↔non-text. We perform the evaluation on two scenarios and four datasets: image↔image (image-image translation) and image↔label (image classification vs. conditioned image generation). In such case where no explicit structure can be used (e.g., syntactic structure), we take the semantic-semantic `RoI` alignment. In Table 11 and 10 we present the results.

B.2. More Results

We further conduct more evaluations on many other datasets, so as to gain some consistent observations. In Table 12 and 13 we show the results on text↔image and text↔label cases.

	CIFAR-10			CIFAR-100		
	IMG→LB	IMG←LB		IMG→LB	IMG←LB	
	ACC	IS↑	FID↓	ACC	IS↑	FID↓
M1	93.05	8.62	13.53	72.60	9.34	19.63
M3	93.68	9.83	9.80	73.85	13.64	15.72
M4	**94.74**	**10.64**	**7.38**	**74.63**	**14.65**	**13.42**
Δ	**+1.06**	**+0.81**	**-2.42**	**+0.78**	**+1.01**	**-2.30**

Table 10. Image↔Label experiment (IMG→LB: image classification, IMG←LB: conditioned image generation) on CIFAR-10 and CIFAR-100 datasets.

	CelebA-HQ		AFHQ	
	$IMG_A \rightarrow IMG_B$	$IMG_A \leftarrow IMG_B$	$IMG_A \rightarrow IMG_B$	$IMG_A \leftarrow IMG_B$
M1	26.7	32.7	32.4	40.8
M3	20.0	24.6	26.2	29.6
M4	**17.5**	**20.3**	**22.0**	**25.7**
Δ	**-2.5**	**-4.3**	**-4.2**	**-3.9**

Table 11. Image↔Image experiment (image-image translation) on CelebA-HQ and AFHQ datasets. Metrics: FID↓.

	IS↑	FID↓	B-4	MTR
M1	2.7	50.6	43.5	26.8
M3	2.9	47.8	47.0	28.3
M4	**3.3**	**42.9**	**53.7**	**32.4**
Δ	**+0.4**	**-4.9**	**+6.7**	**+4.1**

Table 12. Text↔Image experiment (TXT→IMG, TXT←IMG) on CUB data.

	ACC	B-4	MTR	ACC
M1	89.3	10.5	21.1	76.4
M3	90.4	14.8	24.8	80.5
M4	**92.2**	**16.7**	**28.0**	**88.3**
Δ	**+1.8**	**+1.9**	**+3.2**	**+7.8**

Table 13. Text↔Label experiment (TXT→LB, TXT←LB) on AGnews.